The IUCr Diffraction Data Deposition Working Group Activities

John R. Helliwell

School of Chemistry
University of Manchester
Oxford Road
Manchester M13 9PL, UK

john.helliwell@manchester.ac.uk

Brian McMahon

Research & Development Officer
International Union of Crystallography
5 Abbey Square
Chester CH1 2HU, UK

bm@iucr.org

http://forums.iucr.org/ viewforum.php?f=7

Abstract

It is increasingly important to deposit the raw data from scattering experiments; valuable information gets lost when only structure factors are deposited. Some research centres, *e.g.* synchrotron and neutron facilities, are fully aware of the need to archive raw data. Diamond Light Source, ISIS, the European PaN (Photon and Neutron facilities) and the Australian TARDIS initiatives are exemplars of good practice. Local university data repositories are also being developed, *e.g.* at the University of Manchester, relevant to laboratory X-ray diffraction data archiving. A globally registered digital identifier (such as DOI) for each significant dataset can therefore be obtained and linked to any related publication. A Working Group on the development of standards for the representation of data and associated metadata to permit the routine deposition of such raw data has been launched by the IUCr Executive Committee. Its title is 'Diffraction Data Deposition Working Group of the IUCr'. Its members include representatives from IUCr Commissions, synchrotron facilities, academic laboratories, structural databases, and inter-Scientific Union bodies on data and publications. There is a forum for 'Public input on diffraction data deposition' on the IUCr web site at http://forums.iucr.org, to which all interested parties are invited to contribute.

Mining the data

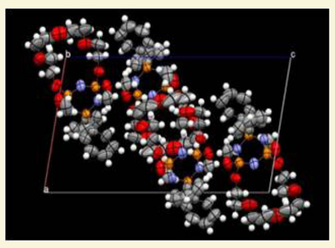
The phrase 'data mining' has often been used to describe the extraction of information from large volumes of available data (*e.g.* performing a search across all the entries in the Cambridge Structural Database or Protein Data Bank). In fact, it also suggests a valuable metaphor for the processing of experimental data through to a published structural model, especially in terms of volume. There is often a huge amount of raw data from an experiment (several gigabytes of image data), similar to the amount of earth that must be excavated in a mining operation. The processed data are much smaller in volume (perhaps only a few megabytes of structure-factor listing), and from this high-grade 'ore' may be refined or extracted the few kilobytes of 'precious metal' that we can identify with the structural coordinates and displacement parameters of the final model. In most mining operations, once processed, the various spent materials are tossed aside as spoil; yet they may contain other minerals or materials of less immediately obvious value, yet interesting in their own right. In the same way, the raw data that are now routinely discarded may contain information about other scientific features that do not interest the primary investigator. But, in principle, they are there to be unearthed by future prospectors. The question is: can we afford such secondary mining processes?

Why publish data?

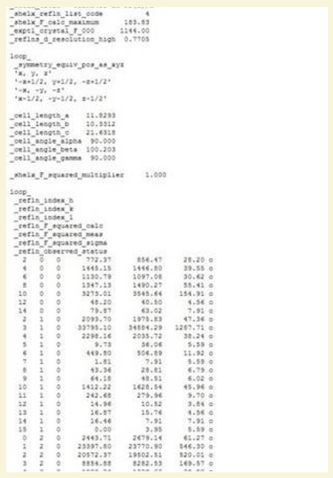
Some reasons:

- To enhance the reproducibility of a scientific experiment
- To verify or support the validity of deductions from an experiment
- To safeguard against error
- To safeguard against fraud
- To allow other scholars to conduct further research based on experiments already conducted
- To allow reanalysis at a later date, especially to extract 'new' science as new techniques are developed
- To provide example materials for teaching and learning
- To provide long-term preservation of experimental results and future access to them
- To permit systematic collection for comparative studies

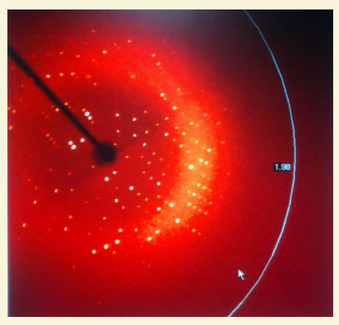
All are – in some measure – 'good' reasons. The question is: how much effort, time and money is it reasonable to spend for the potential benefits that might accrue from publishing particular types of data (raw, processed, derived)?



Derived data are those summarizing the refined structural model.



Processed data are the structure factors or Rietveld profiles reduced by standard techniques from the raw experimental output.



Raw data are those generated by the experimental apparatus.

IUCr journal policy

(1) Derived data

For crystal/molecular structures with small unit cell

- Atomic coordinates, anisotropic displacement parameters, molecular geometry and intermolecular contacts
- Experimental parameters, unit-cell dimensions, space group information
- Reference and modulated structure subsystems for aperiodic composite structures

must be supplied in CIF format as an integral part of article submission and are freely available for download.

For biological macromolecular structures

- Atomic coordinates, anisotropic or isotropic displacement parameters, space group information, secondary structure and information about biological functionality

must be deposited with the Protein Data Bank before or in concert with article publication; the article will link to the PDB deposition using the PDB reference code.

- Relevant experimental parameters, unit-cell dimensions are required as an integral part of article submission and are published within the article.

(2) Processed experimental data

For crystal/molecular structures with small unit cell

- Structure factors
- Rietveld profiles

must be supplied in CIF format as an integral part of article submission and are freely available for download. *SHELXL* instruction files are also required for validation.

For biological macromolecular structures

- Structure factors

must be deposited with the Protein Data Bank before or in concert with article publication; the article will link to the PDB deposition using the PDB reference code.

(3) Primary experimental data

For small-unit-cell crystal/molecular structures and macromolecular structures

IUCr journals have *no* current binding policy regarding publication of diffraction images or similar raw data entities.

However, the journals welcome efforts made to preserve and provide primary experimental data sets, and encouragement is often provided in Notes for Authors. For example, submission guidelines for *Acta Crystallographica Section D* state:

> *Fibre data should contain appropriate information such as a photograph of the data. As primary diffraction data cannot be satisfactorily extracted from such figures, the basic digital diffraction data should be deposited.*
>
> *Authors are encouraged to make arrangements for the diffraction data images for their structure to be archived and available on request.*
>
> *For articles that present the results of powder diffraction profile fitting or refinement (Rietveld) methods, the primary diffraction data, i.e. the numerical intensity of each measured point on the profile as a function of scattering angle, should be deposited.*

Furthermore, many IUCr Commissions are interested in the possibility of establishing community practices for the orderly retention and referencing of such data sets, and the IUCr would like to see such data sets become part of the routine record of scientific research in the future, to the extent that this proves feasible and cost-effective.

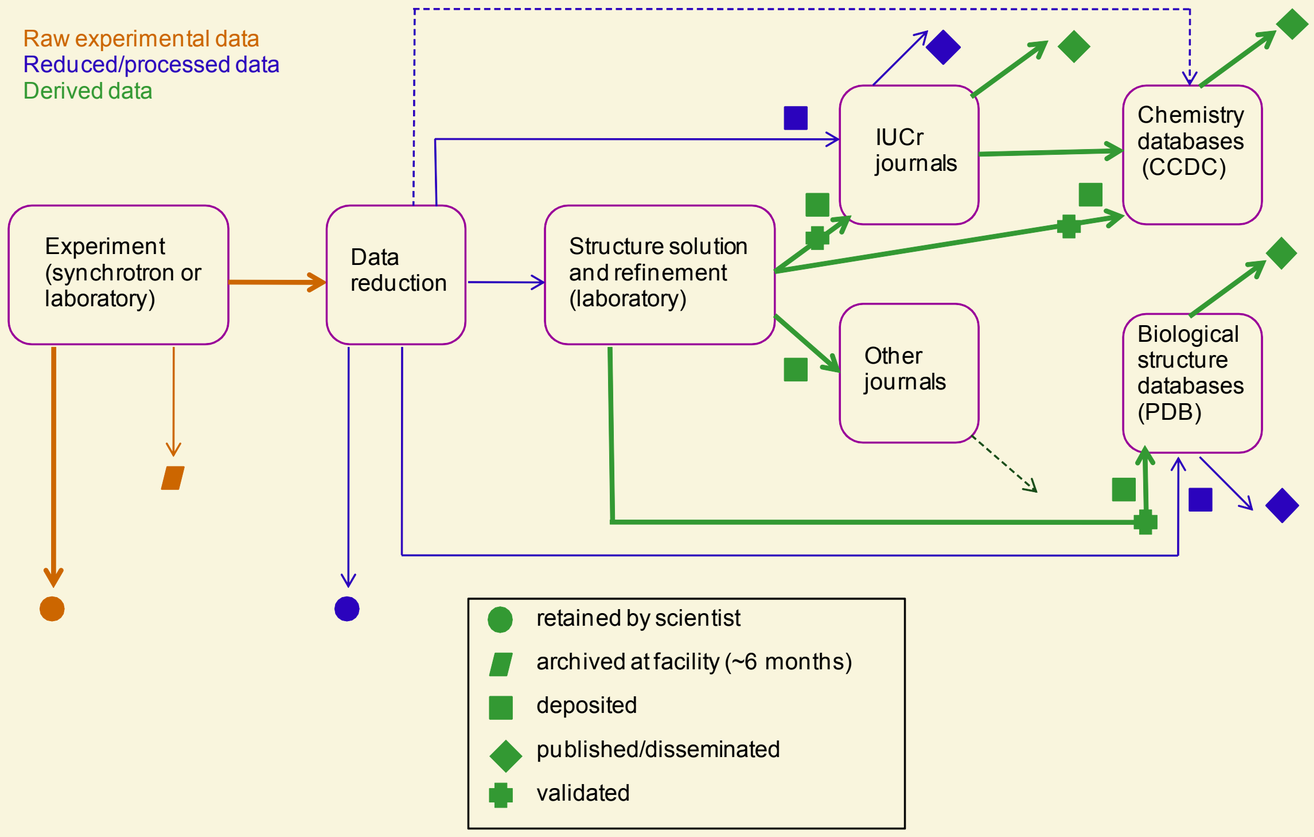
Policy of funding bodies

Increasingly, funding agencies are requesting or requiring data management policies (including provision for retention and access) to be taken into account when awarding grants. See *e.g.* the Research Councils UK Common Principles on Data Policy (http://www.rcuk.ac.uk/research/Pages/DataPolicy.aspx) and the Digital Curation Centre overview of funding policies in the UK (http://www.dcc.ac.uk/resources/policy-and-legal/overview-funders-data-policies).

See also http://forums.iucr.org/viewtopic.php?f=21&t=58 for discussion on policies relevant to crystallography in other countries.

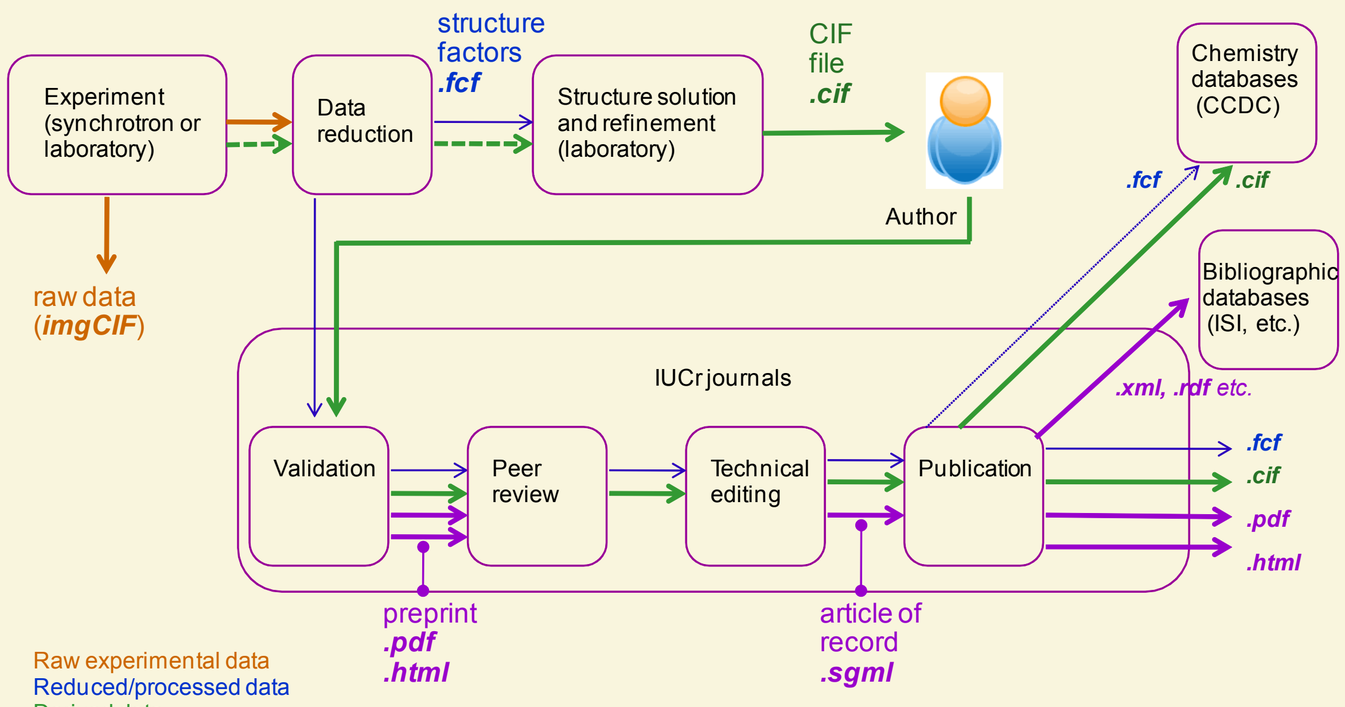
Data flow in crystallography

This schematic indicates the flow of data from experiment, through analysis, publication and deposition of structural results, for a typical single-crystal X-ray structure determination. The relationship between journals and curated databases differs between small-molecule and macromolecular communities. But there are broadly parallel flows, and owing to standard Web-based information access and delivery systems, publishers, databases, laboratories and large-scale facilities all have the potential of working together to create a distributed information ecosystem.



Publication flow in IUCr journals

This schematic shows the information flow associated with publication of crystal structure reports in IUCr journals. The data associated with the publication are handled as outlined in the upper figure, but here we emphasise the role of the author, who interpolates the textual content of a paper directly into a CIF (in the case of *Acta Crystallographica Sections C* and *E*), the value added to the publication through peer review (including the semi-automated analysis and validation performed by the *checkCIF* service), and the multiple format translations through which the data and associated comment comprising an article must pass during the technical editing and publishing processes. It is important that many of the format interconversions can occur losslessly (*i.e.* without decay of information content). In practice, subsets of the information are filtered out during various stages of processing. The article is usually delivered as a HTML or PDF document, with little numerical data retained (except to the extent that it is relevant to the scientific discussion), and less semantic markup than in the input files (thus, PDF is largely an end-stage presentational medium rather than a working format), while the numerical data in the input CIF and structure factor files are usually served as supplementary files.



Working group members

- Steve Androulakis *Representative of TARDIS (Australian repositories for diffraction images)*
- Sol Gruner *Diffuse scattering specialist and Synchrotron Radiation Facility Director*
- John R. Helliwell, **Chair** *IUCr Representative to ICSTI; Chair, IUCr Commission on Journals 1996-2005*
- Loes Kroon-Battenburg *Data processing software developer and user*
- Brian McMahon *IUCr Representative to CODATA*
- Tom Terwilliger *Representative of IUCr Commission on Biological Crystallography*
- John Westbrook *Representative of wwPDB (Worldwide Protein Data Bank)*
- Hans-Josef Weyer *Synchrotron Radiation and Neutron Facility user*

Consultants

- Alun Ashton *Diamond Light Source*
- Herbert Bernstein *Head, imgCIF Dictionary Maintenance Group and member of COMCIFS*
- Frances Bernstein *Observer on data deposition policies*
- Gerard Bricogne *Active software and methods developer*
- Bernhard Rupp *Macromolecular crystallographer*

References and further reading

Research data management policy: UK Research Data Service, *The data imperative: Managing the UK's research data for future use* (2009). Available from:
http://www.ukrds.ac.uk/resources/download/id/14

Research data management policy: e-IRG Data Management Task Force. *Report on Data Management* (2009). Available from:
http://www.e-irg.eu/images/stories/publ/task_force_reports/dmtfjointreport.pdf

Research data management policy: High Level Expert Group on Scientific Data *Riding the wave - How Europe can gain from the rising tide of scientific data* (2010). Available from:
http://ec.europa.eu/information_society/newsroom/cf/itemlongdetail.cfm?item_id=6204

Domain repository solutions: S. Androulakis, J. Schmidberger, M. A. Bate, R. DeGori, A. Beitz, C. Keong, B. Cameron, S. McGowan, C. J. Porter, A. Harrison, J. Hunter, J. L. Martin, B. Kobe, R. C. J. Dobson, M. W. Parker, J. C. Whisstock, J. Gray, A. Treloar, D. Groenewegen, N. Dickson and A. M. Buckle *Federated repositories of X-ray diffraction images* (2008). *Acta Cryst.* D**64**, 810-814 [doi:10.1107/S0907444908015540]

Infrastructure architecture: Photon and Neutron Open Data Infrastructure (2012) http://pan-data.eu

Persistent identifiers for data: DataCite (2009). http://www.datacite.org/